Seattle

Projet_04 :

Anticipez les besoins en consommation de bâtiments.

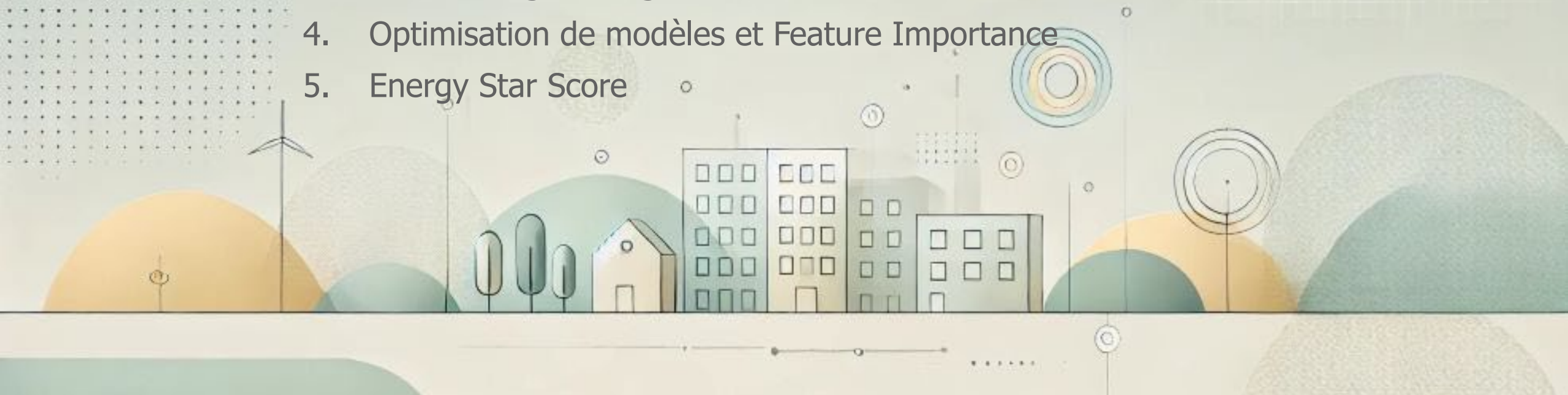
Jérôme LE GAL
Etudiant OpenClassRooms – parcours Data Scientist

Le 12/08/2024

- **Contexte du projet:**
  - Mission pour la ville de Seattle
  - Prédire les **émissions de CO2** et la **consommation totale d'énergie** de bâtiments non destinés à l'habitation.

- **Objectifs de l'analyse :**

1. Nettoyage des données et analyse exploratoire
2. Tester différents modèles de prédiction et évaluer leurs performances
3. Feature Engineering
4. Optimisation de modèles et Feature Importance
5. Energy Star Score



Description du jeu de données

Contenu :

- Données géographiques
- Superficies
- Usages
- Energies (type, consommations,…)

- *Source : 2016 Building Energy Benchmarking*

https://www.seattle.gov/tech/reports-and-data/open-data

Choix des targets

Target 1
SiteEnergyUse(kBtu)

Target 2
TotalGHGEmissions

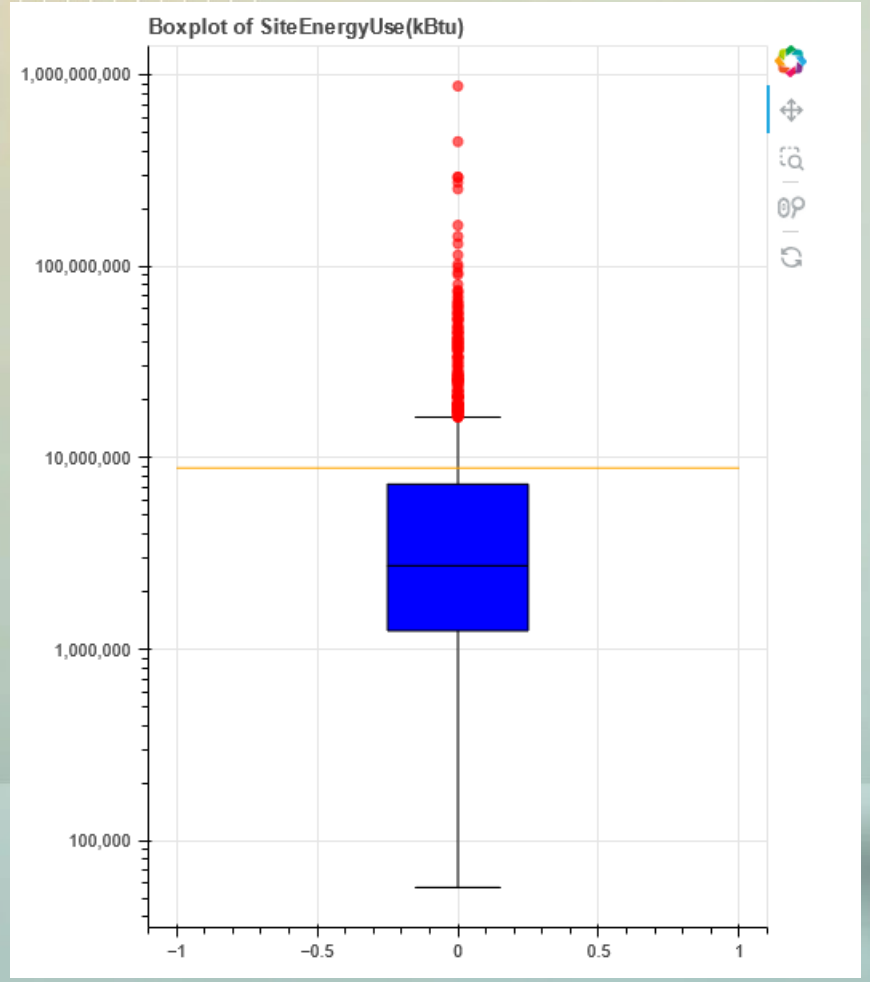
($kgCO_2e/ft^2$)

Analyse exploratoire des données

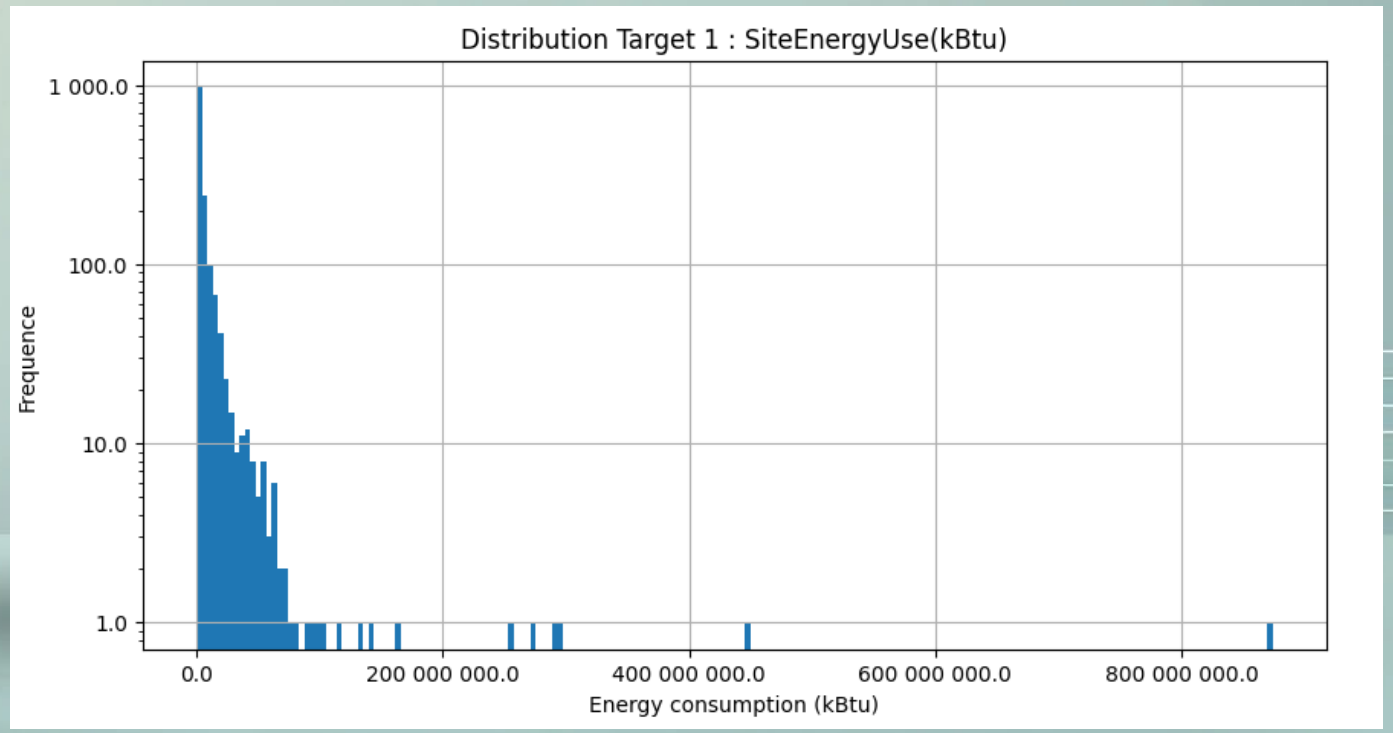
Target 1 :

- SiteEnergyUse (kBtu) :

Mean	8 860 058
Std	31 305 680
Min	57 133
Max	873 923 700



- Sélection des features pertinentes (12)
- Traitement des manquants et doublons
- Traitement des outliers

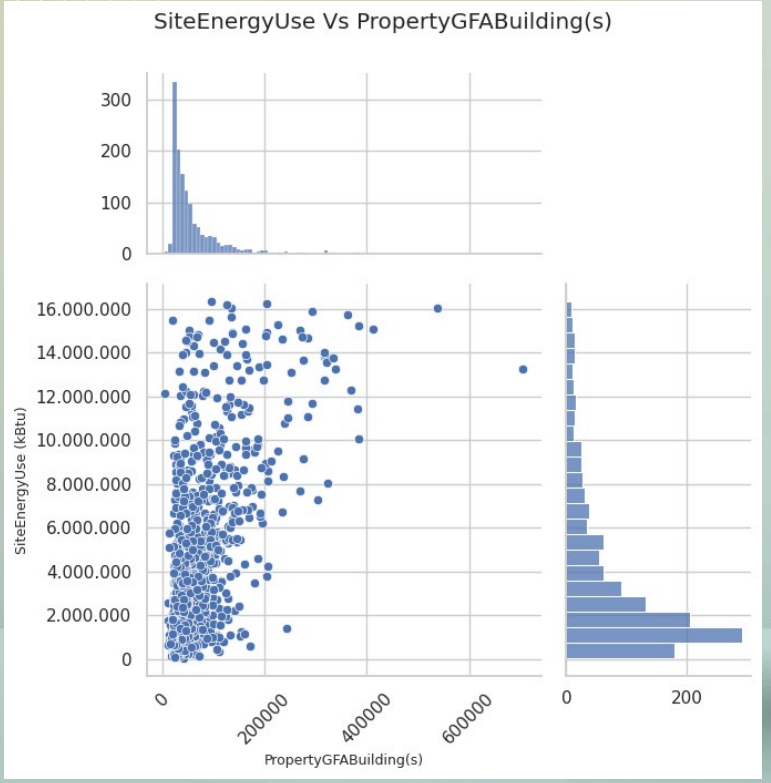


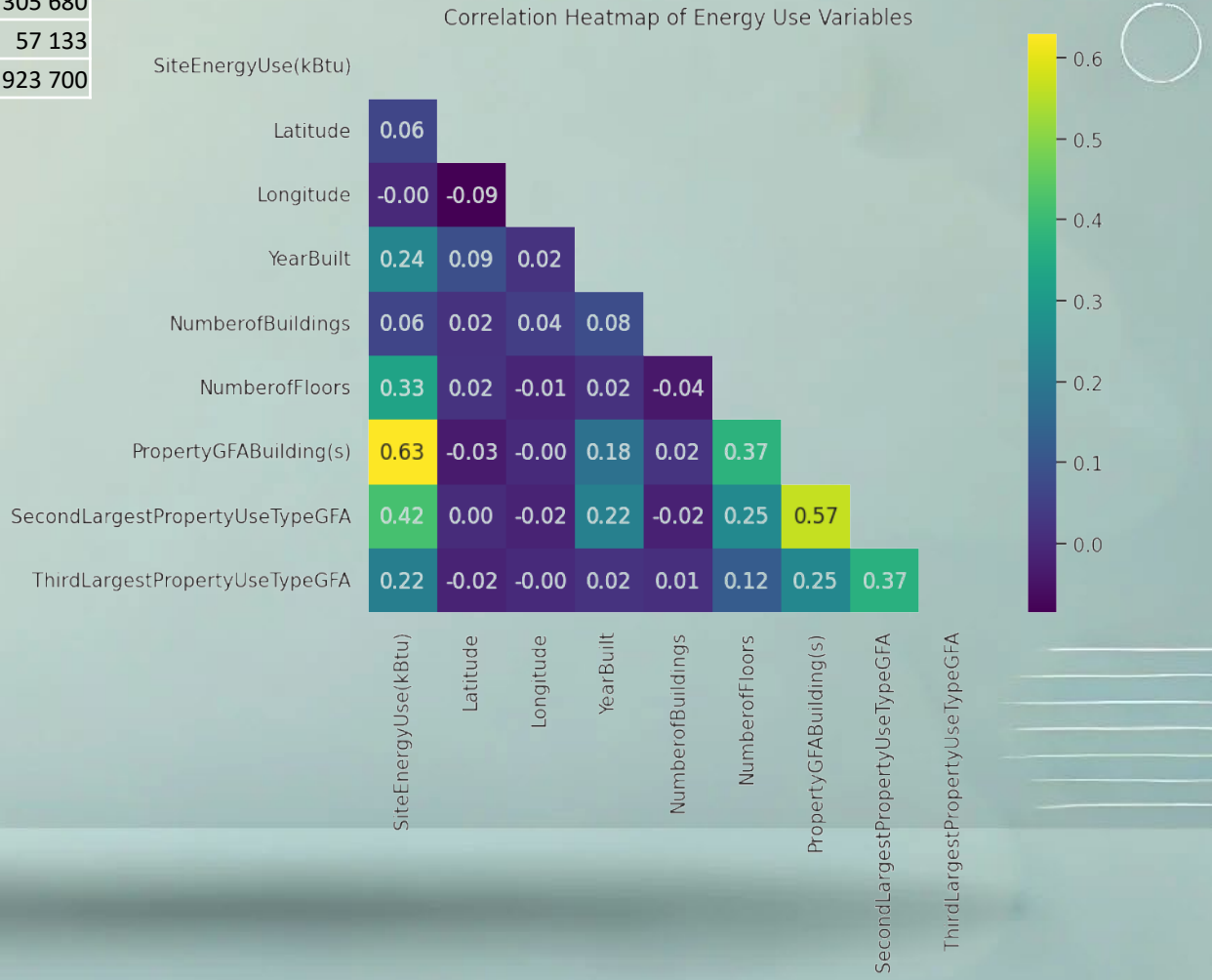
Analyse exploratoire des données

Target 1 :

- SiteEnergyUse (kBtu) :

Mean	8 860 058
Std	31 305 680
Min	57 133
max	873 923 700





Feature Engineering

Target 1 :

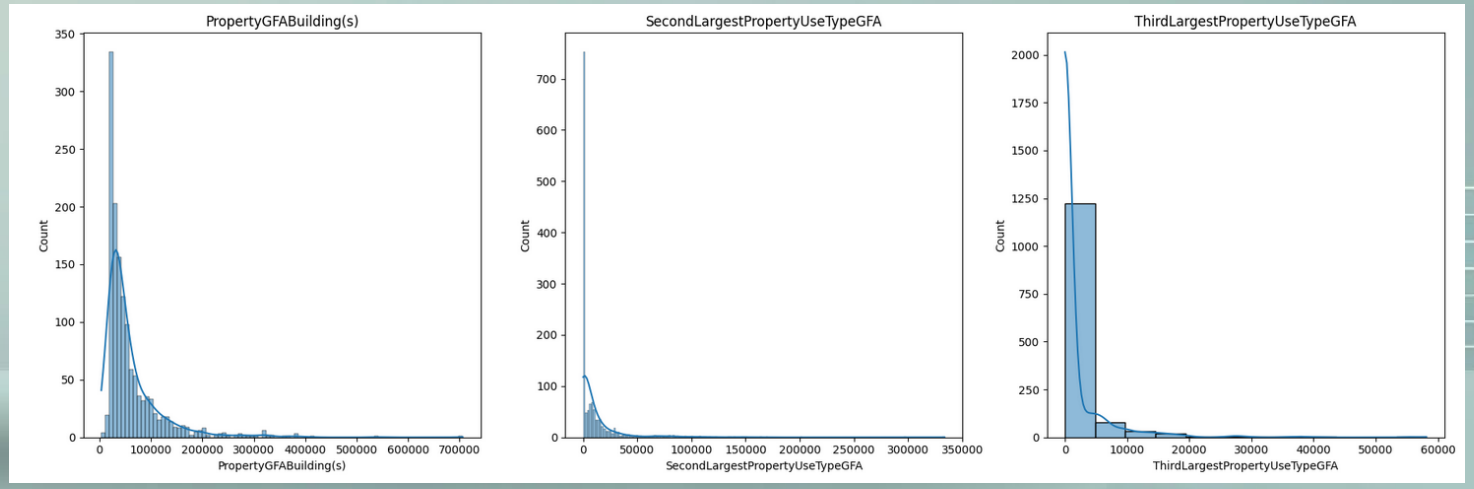
1 - Créations de nouvelles variables :

- Âge des bâtiments
- Ratio électricité consommée
- Ratio gaz naturel consommé
- Ratio Vapeur consommée
- Nombre d'usage

2 - Encodage des variables qualitatives :

- OneHotEncoder

3 - Transformation logarithmique :



4 – StandardScaler() :

- Centrer / Réduire

Modélisations et validation

- **Choix de algorithmes :** LinearRegression, Bagging Regressor, Random Forest Regressor, AdaBoost Regressor, Stacking Regressor, Gradient Boosting Regressor
- **Validation croisée :** Utilisation de la validation croisée pour évaluer les performances
- **Optimisation des hyperparamètres :** GridSearchCV et RandomSearchCV (comparaison des 2 méthodes)
- **Evaluation des performances :** Mesures (R^2, RMSE, MAE)

Résultats et Comparaisons

- **Performances des modèles :**

	Target 1 - SiteEnergyUse	
Mesure : R^2 ajusté	Train / Test	CV - Test set
Linear Regression	0,679 / 0,615	0,358 | 0,555 | 0,092 | 0,282 | 0,581
Bagging Regressor	0,947 / 0,671	0,362 | 0,598 | 0,454 | 0,415 | 0,657
Random Forest Regressor	0,854 / 0,669	0,372 | 0,593 | 0,404 | 0,394 | 0,668
Adaboost Regressor	0,570 / 0,566	0,261 | 0,471 | 0,240 | 0,238 | 0,589
Stacking Regressor	0,641 / 0,642	0,251 | 0,557 | 0,186 | 0,385 | 0,476
Gradient Boosting Regressor	0,866 / 0,692	0,403 | 0,620 | 0,475 | 0,452 | 0,682

- Optimisation des hyperparamètres :

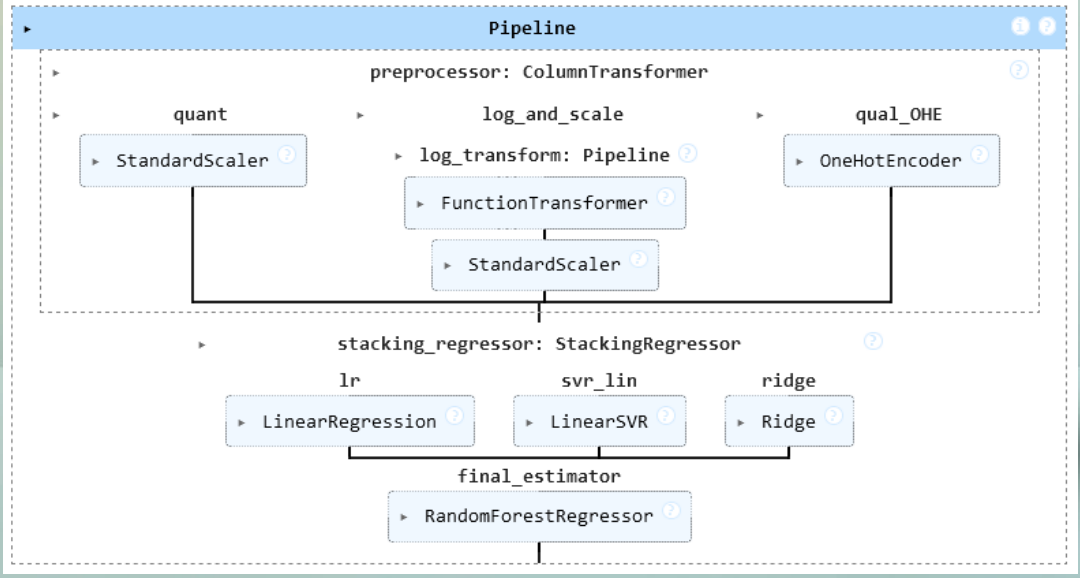
- Interprétations des résultats
- Avantages et inconvénients

Choix et Justifications

- **Choix du modèle et hyperparamètres :**

	Target 1 - SiteEnergyUse	
Mesure : R^2 ajusté	Train / Test	CV - Test set
Linear Regression	0,679 / 0,615	0,358 | 0,555 | 0,092 | 0,282 | 0,581
Bagging Regressor	0,947 / 0,671	0,362 | 0,598 | 0,454 | 0,415 | 0,657
Random Forest Regressor	0,854 / 0,669	0,372 | 0,593 | 0,404 | 0,394 | 0,668
Adaboost Regressor	0,570 / 0,566	0,261 | 0,471 | 0,240 | 0,238 | 0,589
Stacking Regressor	0,641 / 0,642	0,251 | 0,557 | 0,186 | 0,385 | 0,476
Gradient Boosting Regressor	0,866 / 0,692	0,403 | 0,620 | 0,475 | 0,452 | 0,682

StackingRegressor

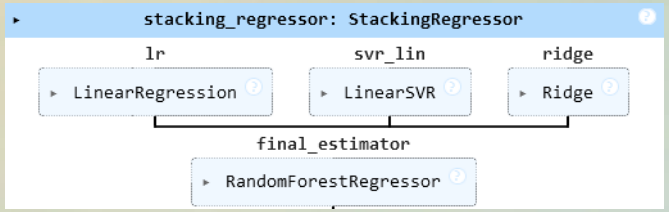


Hyperparamètres :

- LinearRegression
- LinearSVR : C=1
- Ridge : alpha=1
- RandomForestRegressor : n_estimators=100

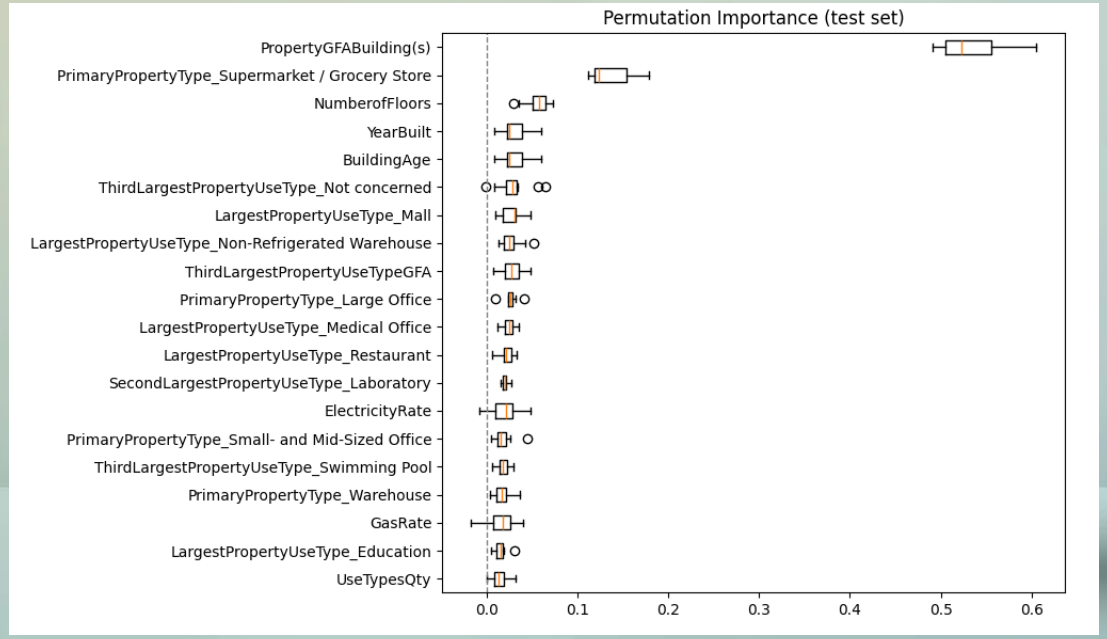
Feature Importance – Target 1

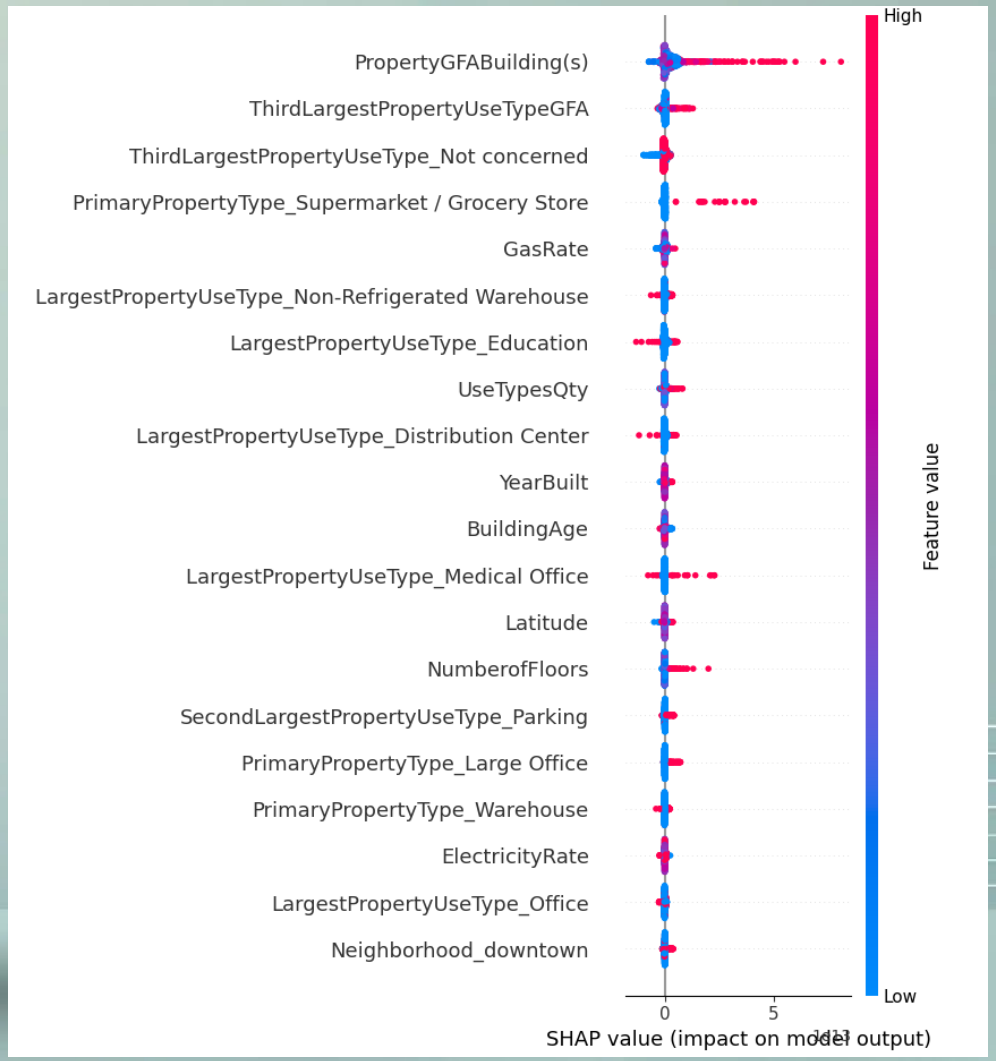
SiteEnergyUse



- **Analyse Globale :**

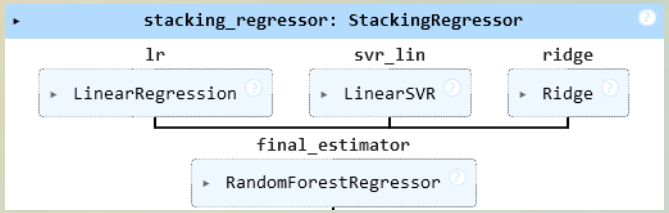
Les 20 features les plus influentes, « Permutation Importance » et méthode SHAP.





Feature Importance – Target 1

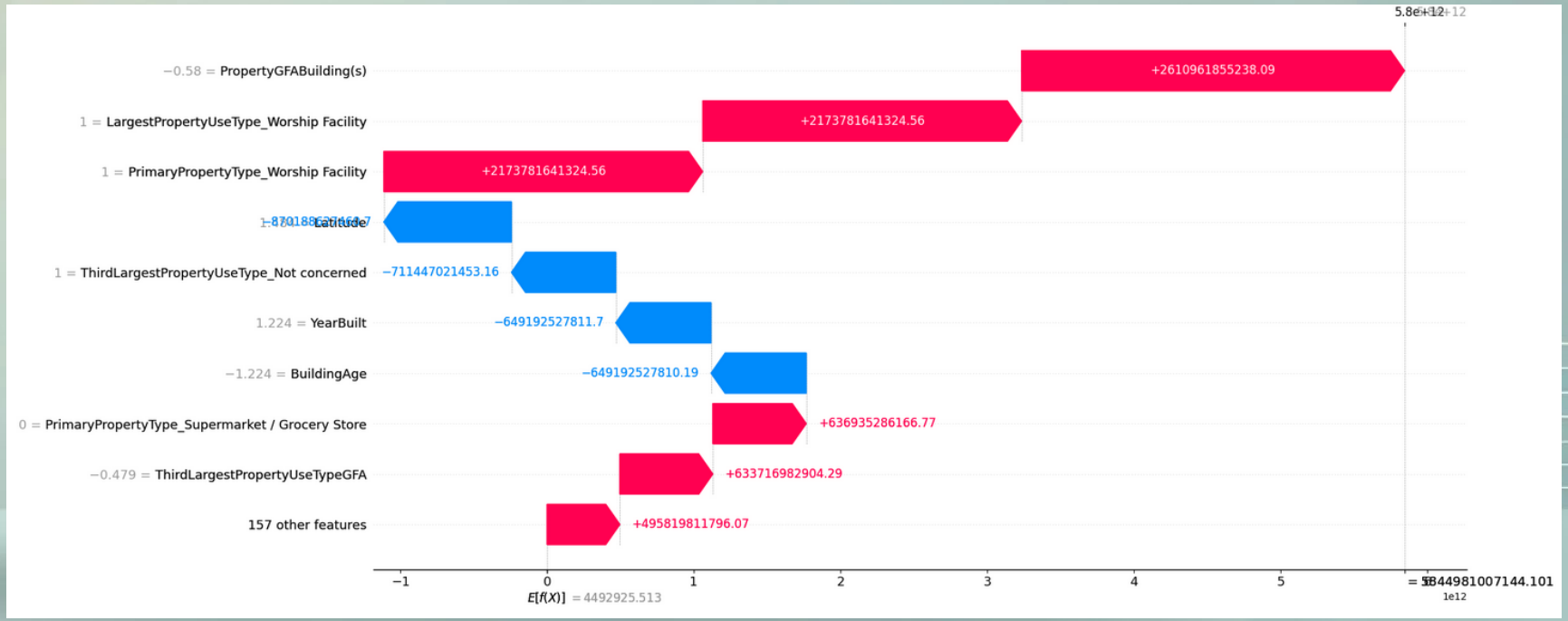
SiteEnergyUse



- **Analyse Locale :**

Latitude	47.68752
Longitude	-122.29852
YearBuilt	2000.0
NumberofBuildings	1.0
NumberofFloors	2.0
BuildingAge	16.0
ElectricityRate	0.629756
GasRate	0.370244
SteamRate	0.0
UseTypesQty	1.0
PropertyGFABuilding(s)	31386.0
SecondLargestPropertyUseTypeGFA	0.0
ThirdLargestPropertyUseTypeGFA	0.0
BuildingType	NonResidential
Neighborhood	northeast
PrimaryPropertyType	Worship Facility
LargestPropertyUseType	Worship Facility
SecondLargestPropertyUseType	Not concerned
ThirdLargestPropertyUseType	Not concerned

Exemple avec SHAP

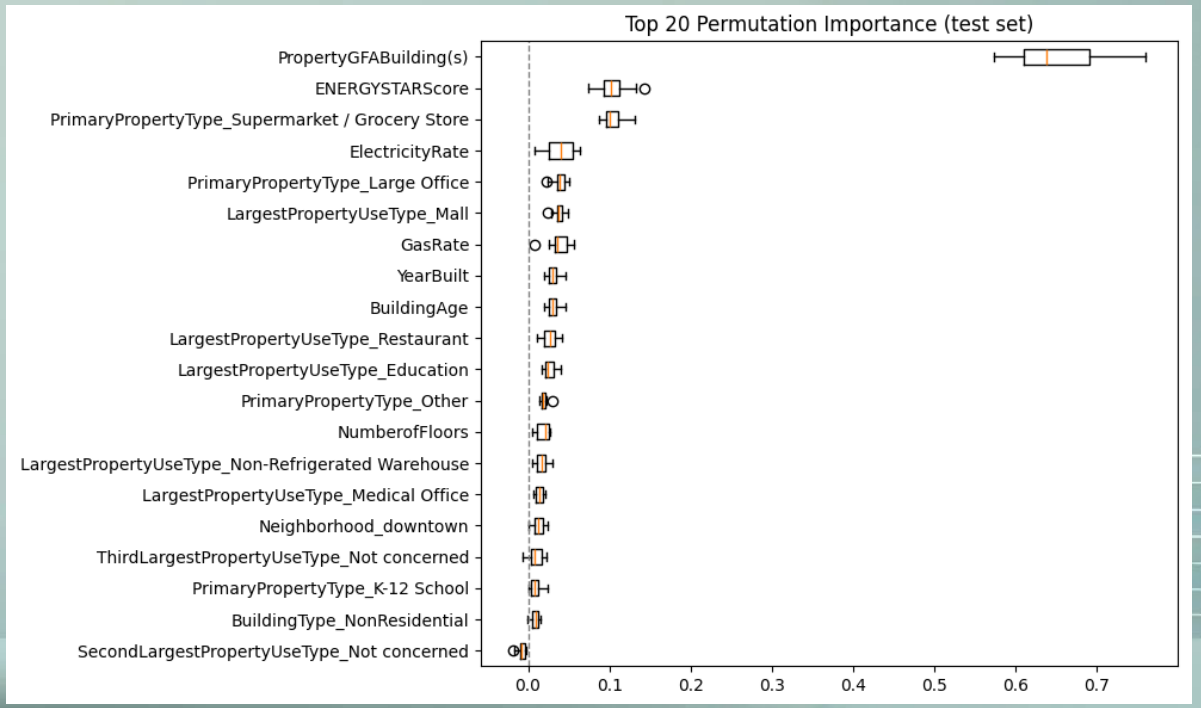


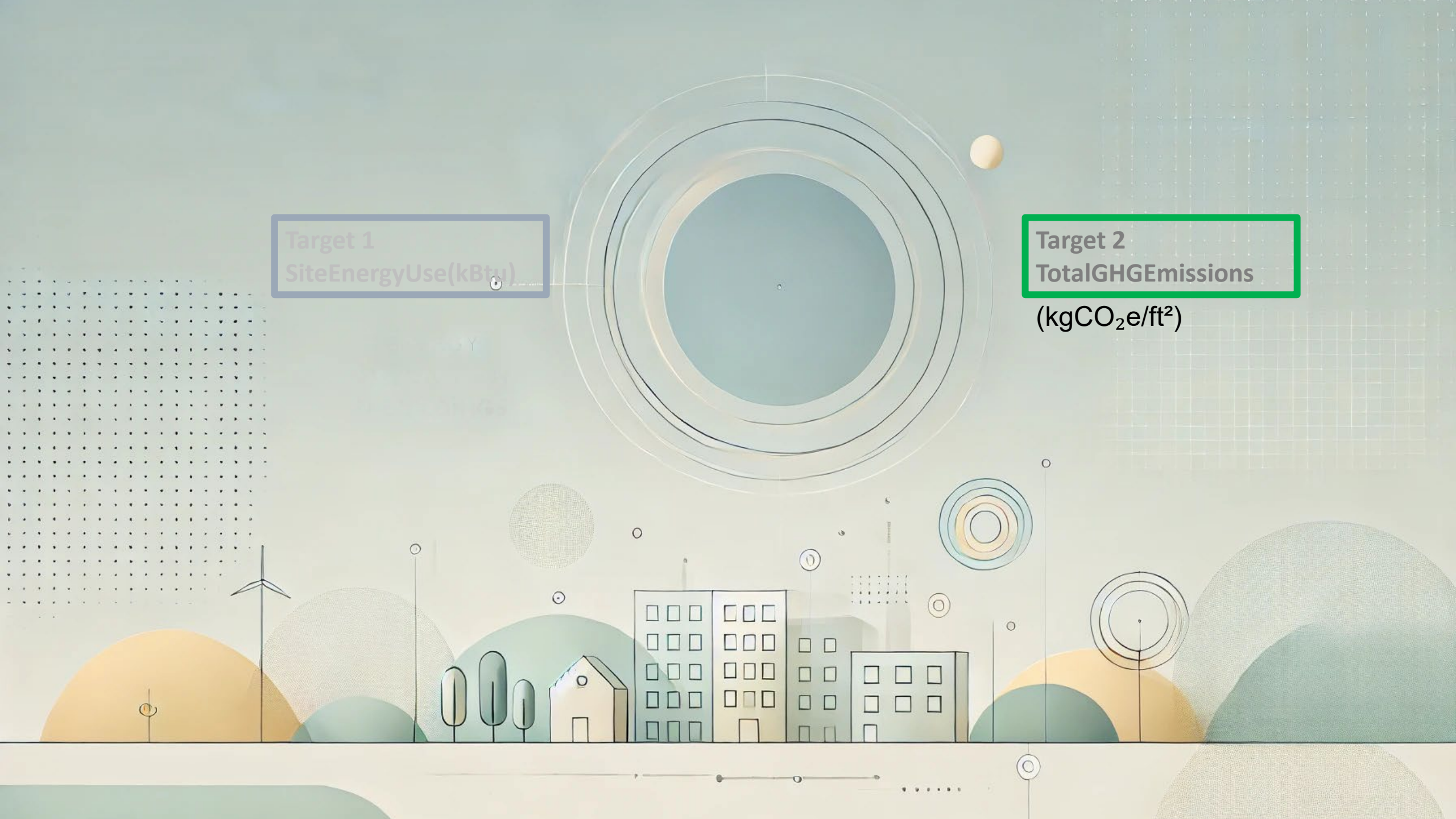
Influence de l'EnergyStarScore

Target 1 :

- **Feature « ENERGYSTARScore » :**
  - **35% de manquants traités par imputation KNN (calcul biais)**

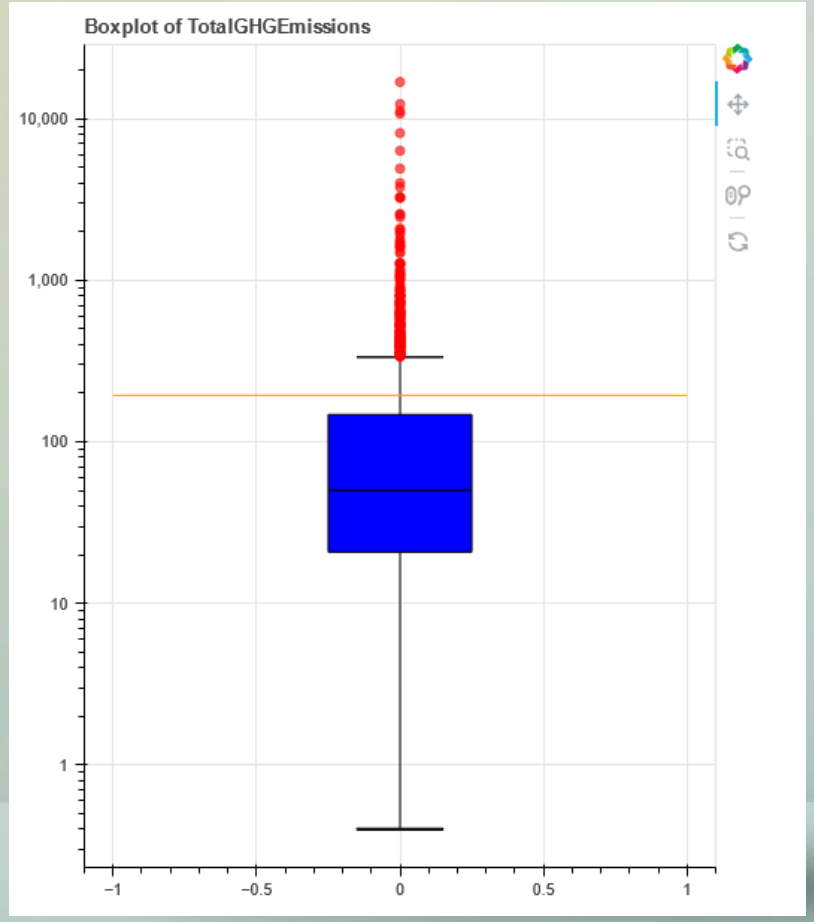
Metric	Avant	Après
R^2_train	0.6423	0.7324
R^2_train ajusté	0.6360	0.7274
R^2_test	0.6483	0.7216
R^2_test ajusté	0.6421	0.7166
Cross_Validation (moyenne)	0.3315	0.4302
Cross_Validation (détail)	[0.2002 0.4901 0.2753 0.2779 0.4142]	[0.2366 0.6068 0.3200 0.3945 0.5931]




Target 1
SiteEnergyUse(kBtu)
Target 2
TotalGHGEmissions
(kgCO2e/ft²)

Analyse exploratoire des données

Target 2 :



- TotalGHGEmissions : $kgCO_2e/ft^2$

Mean	193,61
Std	779,11
Min	-0,80
Max	16 870,98

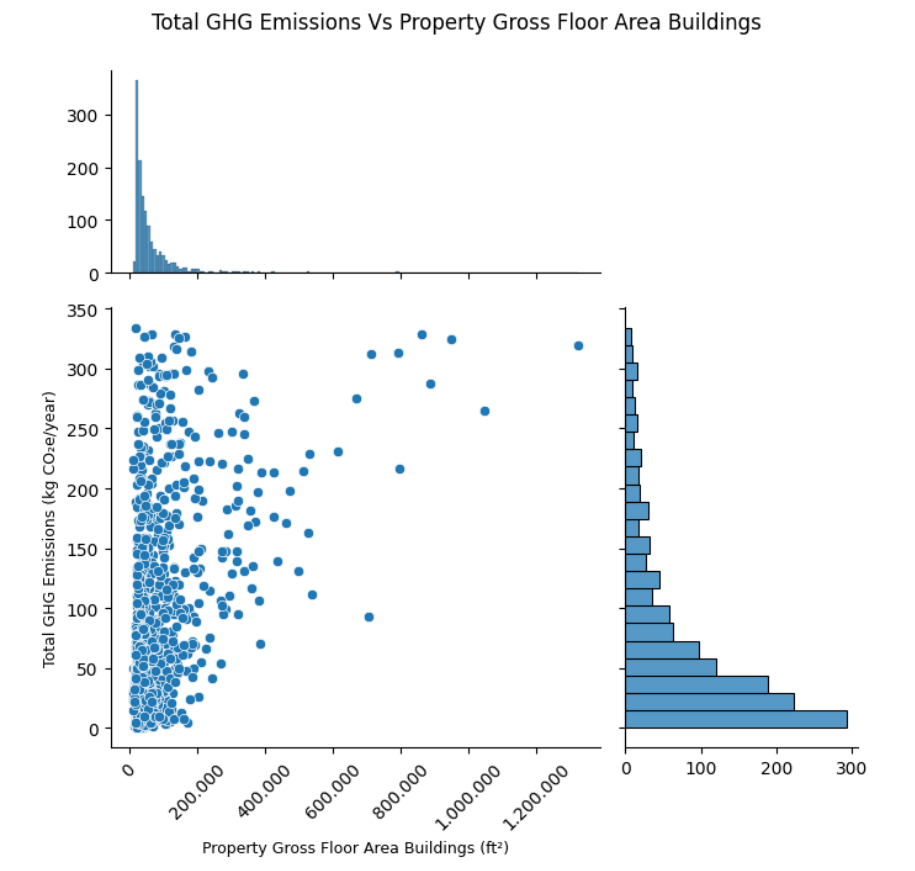
- Sélection des features pertinentes
- Traitement des manquants et doublons
- Traitement des outliers

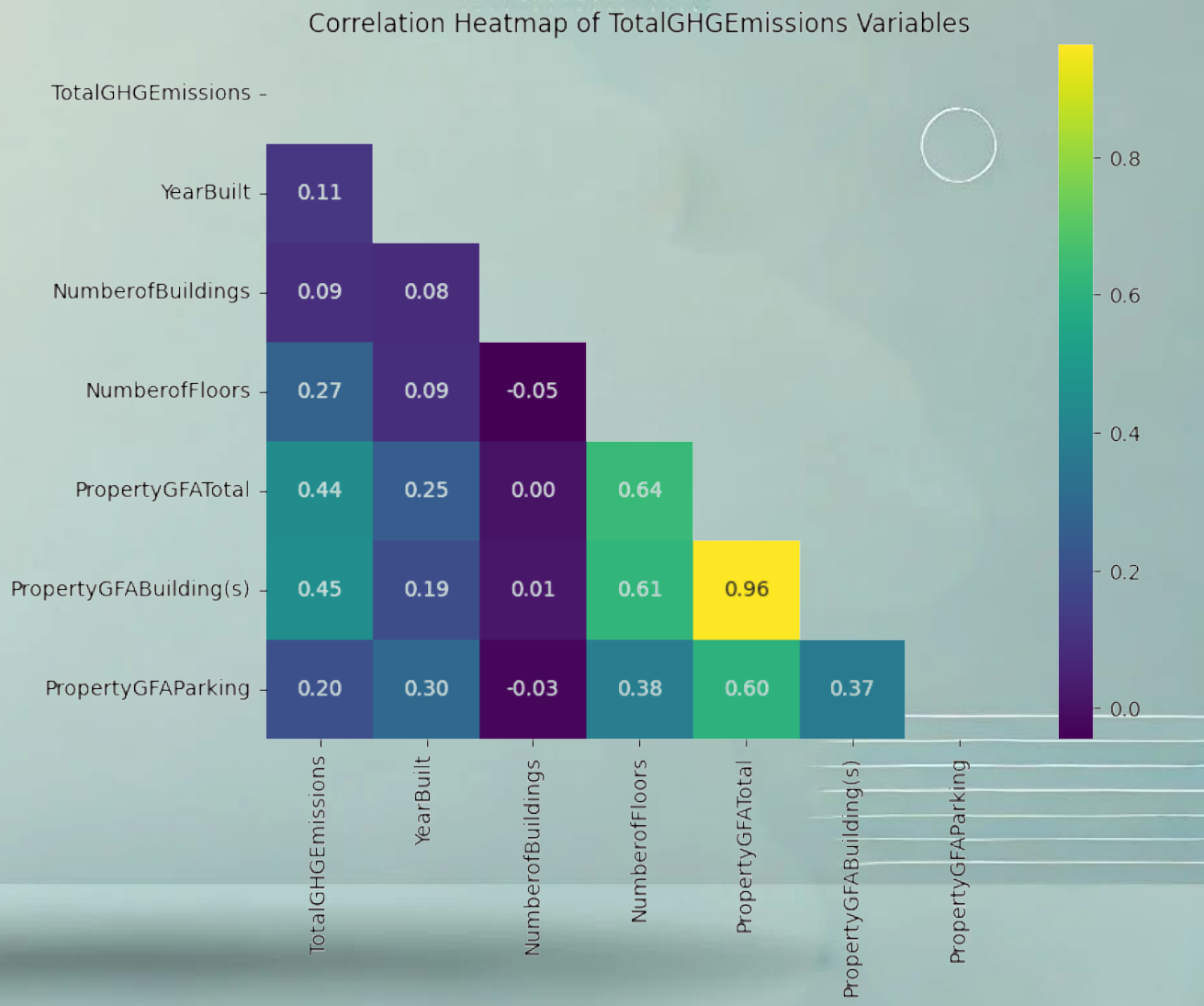
Analyse exploratoire des données

Target 2 :

- TotalGHGEmissions : $kgCO_2/ft^2$

Mean	193,61
Std	779,11
Min	-0,80
max	16 870,98





Feature Engineering

Target 2 :

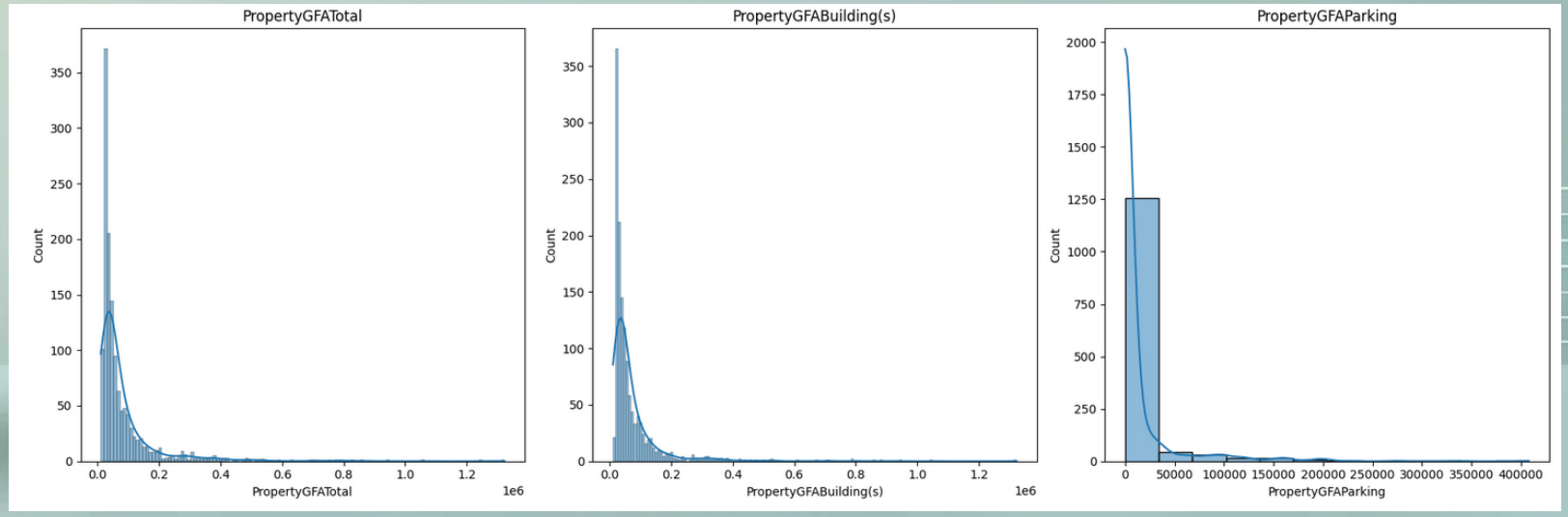
1 - Créations de nouvelles variables :

- Âge des bâtiments
- Utilisation de l'électricité *(variable binaire)*
- Utilisation du gaz naturel *(variable binaire)*
- Utilisation de la vapeur *(variable binaire)*
- Ratio surface bâtiments
- Ratio surface des parkings

2 - Encodage des variables qualitatives :

- OneHotEncoder

3 - Transformation logarithmique :



4 – StandardScaler() :

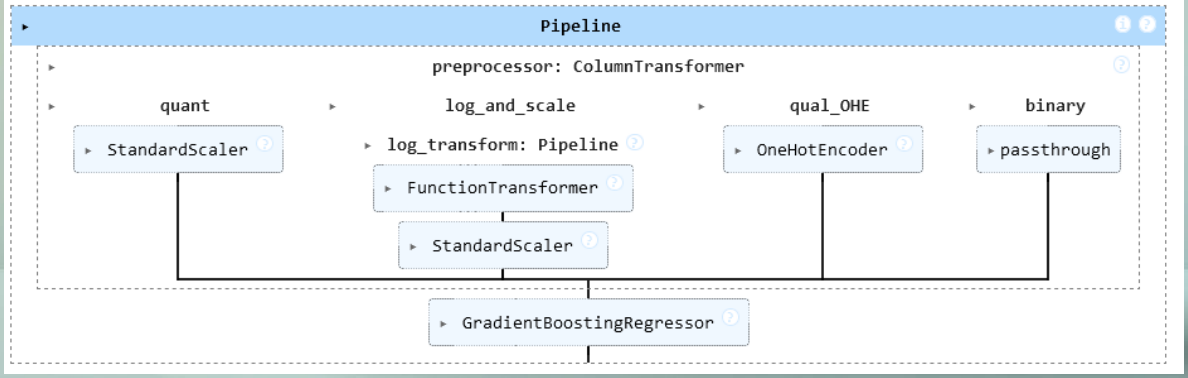
- Centrer / Réduire

Modélisations et résultats

- **Choix du modèle :**

Mesure : R^2 ajusté	Target 2 - TotalGHGEmissions	
	Train / Test	CV - Test set
Linear Regression	0,601 / 0,512	0,368 | 0,413 | 0,340 | 0,353 | 0,313
Bagging Regressor	0,926 / 0,590	0,321 | 0,432 | 0,415 | 0,432 | 0,376
Random Forest Regressor	0,808 / 0,581	0,320 | 0,428 | 0,436 | 0,448 | 0,400
Adaboost Regressor	0,415 / 0,443	0,210 | 0,287 | 0,307 | 0,243 | 0,310
Stacking Regressor	0,534 / 0,519	0,322 | 0,339 | 0,274 | 0,384 | 0,336
Gradient Boosting Regressor	0,730 / 0,599	0,417 | 0,432 | 0,453 | 0,472 | 0,404

Gradient Boosting Regressor



Hyperparamètres :

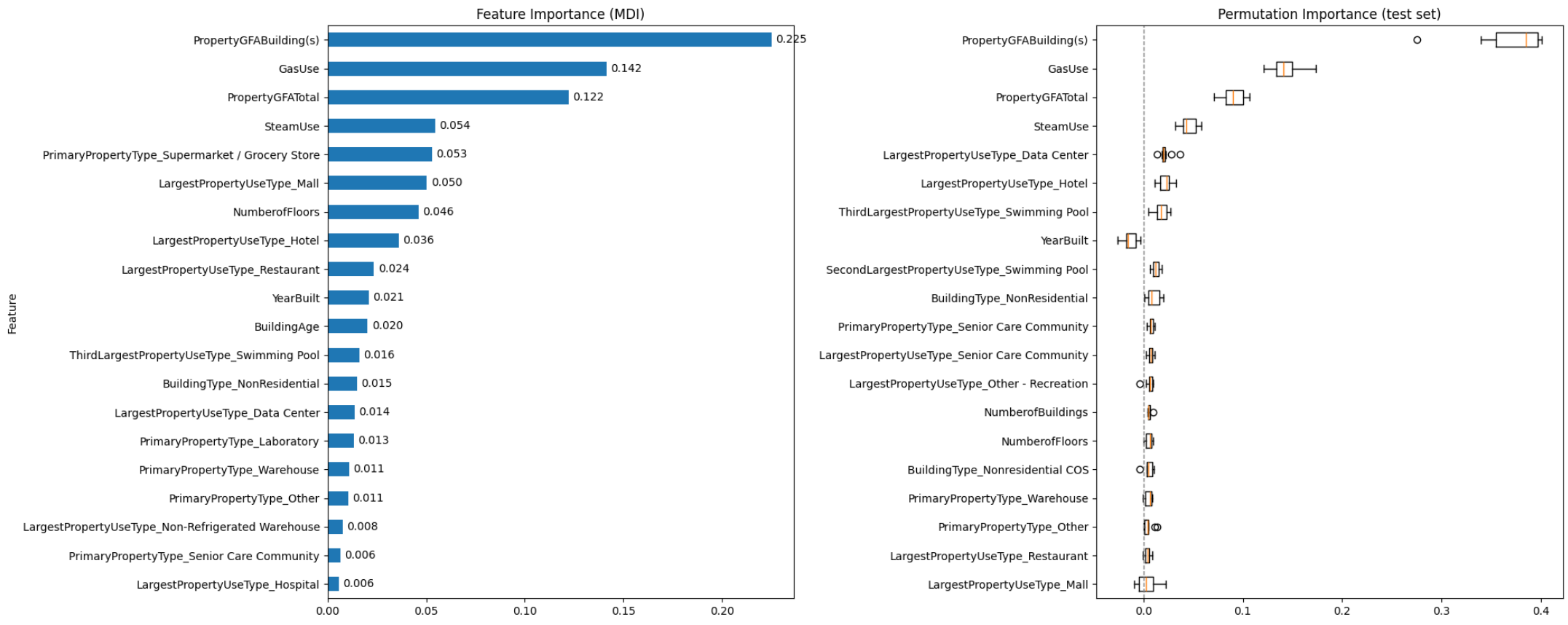
- n_estimators=200
- learning_rate=0.05
- max_depth=3
- subsample=0.9

Feature Importance – Target 2

TotalGHGEmissions

- **Analyse Globale :**

Les 20 features les plus influentes.



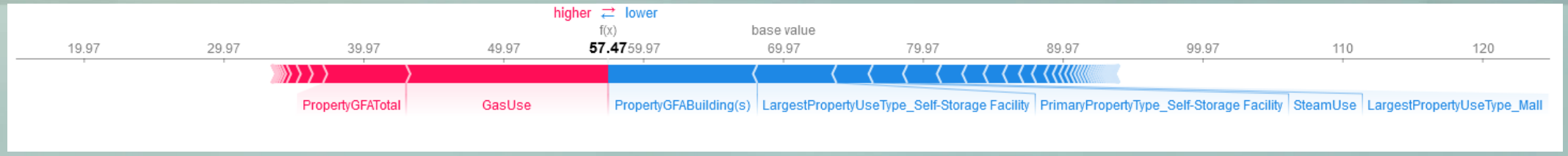
Feature Importance – Target 2

TotalGHGEmissions

- **Analyse Locale :**

Exemple avec SHAP

YearBuilt	2006.0
NumberofBuildings	1.0
NumberofFloors	6.0
BuildingAge	10.0
PropertyGFABuildingRatio	0.851969
PropertyGFAParkingRatio	0.148031
PropertyGFATotal	48179.0
PropertyGFABuilding(s)	41047.0
PropertyGFAParking	7132.0
BuildingType	NonResidential
LargestPropertyUseType	Self-Storage Facility
PrimaryPropertyType	Self-Storage Facility
SecondLargestPropertyUseType	Not concerned
ThirdLargestPropertyUseType	Not concerned
ElectricityUse	1.0
GasUse	1.0
SteamUse	0.0

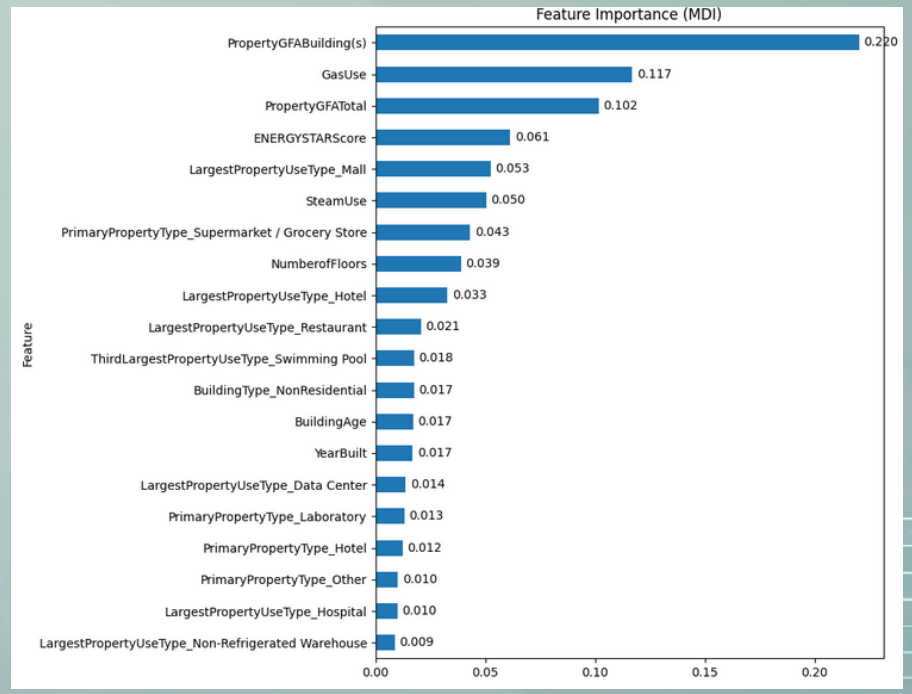


Influence de l'EnergyStarScore

Target 2 :

- **Feature « ENERGYSTARScore » :**
  - **35% de manquants traités par imputation KNN (calcul biais)**

Metric	Avant	Après
R^2_train	0.7339	0.7529
R^2_train ajusté	0.7297	0.7488
R^2_test	0.6049	0.6500
R^2_test ajusté	0.5987	0.6445
Cross_Validation (moyenne)	0.4356	0.4568
Cross_Validation (détail)	[0.4177, 0.4320, 0.4527, 0.4718, 0.4037]	[0.4366, 0.4743, 0.4995, 0.4842, 0.3875]



Conclusion :

- En fonction des besoins de précision :
  - Réalisation de prédictions par tranches (ex: conso entre 1000 et 1500 kBtu)

- Modèles encore perfectibles :
  - Ajout de données d'entrée
  - Amélioration du Feature Engineering (suite ajout Energy Star Score)

Des questions ?